Integration of Biological Data From Web Resources : Management of Multiple Answers Through Metadata Retrieval

Marie-Dominique Devignes and Malika Smaïl

UMR LORIA 7503, CNRS-University Henri Poincaré, BP 239, 54506 Vandoeuvre-lès-Nancy, France.

Received line

ABSTRACT

Biological data retrieval from web resources often necessitates multi-step access to multiple information sources. User-designed scenarios are exploited by a generic application (Xcollect) that allows users to execute them and to store the collected data in a document. Multiple answers are obtained at given steps of the scenario when several resources are queried with same purpose. We address the problem of managing such multiple answers when retrieving functional annotations of genes. Relevant quality metadata for sources and source entries have been listed in view of sorting the answers. Work in progress deals with semantic integration based on domain ontologies.

Contact: {devignes,malika}@loria.fr

INTRODUCTION

Exploiting at best all the mass of biological information stored in the numerous and heterogeneous public data sources is the next challenge in bio-informatics. Various functionalities are proposed by resource providers in terms of databank retrieval or analysis tools. Integrated systems exist that offer unified access to heterogeneous sources and resources. Mediation architectures allow in certain case-studies automatic processing of complex queries. Existing solutions, some of which will be reviewed in section 1, address specific categories of problems but tend to lack flexibility when dealing with any biological question.

Our work is based on the distinction between two levels of problem analysis : first the design of a retrieval scenario involving relevant sources, second, the enactment of this scenario to collect and integrate desired data. At the first level, most users wish to keep the control of scenario design, so that their personal preferences and expertise about sources could be taken into account. A generic model has been produced to allow users describing their scenarios so that an automated solution can be envisaged to deal with the second level of the problem. We present in section 2 the Xcollect application that supports such scenario description as well as its execution. Automation of the data collecting process allows taking into account the frequent changes in source contents by refreshing the data in a time-saving manner.

We then address the management of multiple answers to a given step in the scenario as encountered when retrieving functional annotations for a gene product (section 3). Multiple answers are considered as "homologous" when they correspond to queries formulated with the same intention and to data that share similar semantics. Managing such multiple answers includes (i) sorting them according to user-defined criteria, likely quality criteria, and (ii) integrating them in terms of consistency, discrepancy, precision etc. Taking into account metadata about sources and data is presented here as a mean to sort out the answers and to make decision in case of inconsistency.

Finally we conclude about the perspectives raised by this work with regards to recent development projects around semantic web and web services in bio-informatics.

RELATED WORK ON BIOLOGICAL RESOURCES INTEGRATION

Many available biological sources offer an integrated access to the data thanks to cross-references. For example, hypertext links contained in a SWISS-PROT entry allow hopping to other related data sources (http://us.expasy.org/sprot/). Data retrieval is then more related to browsing than to querying. Unifying the query interface and query language has been implemented by Entrez (http://www.ncbi.nlm.nih.gov/) or SRS (http://srs.ebi.ac.uk) multi-bases systems. In SRS, views can be defined by users to integrate data that are retrieved from several sources. Several services also exist that propose the execution of scenarios composed of chained steps thanks to integrative platforms : task concept at HUSAR bioinformatics (Ernst et al., 2003) ; protocols in BioNavigator from Entigen Corporation (http://www.bionavigator.com), etc. In such situations, the user must deal with predefined choices concerning sources and resources.

Ideally source integration corresponds to data integration in the sense that the distribution of a complex query to various heterogeneous data sources becomes transparent to the user who gets the impression to query a unique source. Approaches such as data warehouses and mediation architectures address this problem. These approaches have in common that the schemas of a collection of relevant data sources have been merged to form a global schema in a common model. Users query this global schema using a high-level query language. In mediation architectures, also referred to as "view integration systems", the system determines what portion of the global query can be answered

by which underlying data source, ships local queries off to the underlying data sources, and combines the answers from the underlying data sources to produce an answer to the global query. Wrappers have to be developed to provide access to remote data sources and to translate/transform the answers into the common representation. The main advantage of these systems is that they preserve source autonomy. Various systems have been developed on the basis of non-materialized views of biological data sources, such as K2/Kleisli, TINet, P/FDM, DiscoveryLink, TAMBIS (Goble et al., 2001), as discussed in Eckman et al. (2001). However access to the systems and covered biological areas are limited.

Data warehouses can be seen as instanciations of the global schema. Data are imported and the global query can then be answered locally. The main advantages of the warehouse solution are performance and the great added value to the data stored in the warehouse (Davidson et al., 2001). The main problem with warehouses is the cost of maintenance. Updating the data in the warehouse is a critical issue. Another problem is that of tracking the origin of a piece of data or annotation. Complexity and costs of maintenance make large data warehouse impractical for small biology laboratories. This system becomes realistic at moderate scale when dealing with a limited set of data sources and/or for projects involving production strength such as large scale annotation projects (for example : GUS, InterPro, BioMolQuest).

XCOLLECT : A GENERIC "USER-ORIENTED" APPROACH

In a previous work a dedicated application (Xmap project) had been developed to deal with a specific biological question (Devignes et al., 2002). A data retrieval process based on a generic scenario model has then been implemented in the Xcollect application. The generic scenario model appears as a succession of steps described in the XML Xcollect scenario_DTD. For each step, following information is specified : (1) source name and location, (2) input formal name and value (inputs include parameters for query construction, as well as relevant data retrieved at any previous step), (3) output formal name and type, (4) patterns necessary to extract the useful data from the returned document (*e.g.* regular expressions).

Structuring the retrieved data also implies a model. In the absence of any existing standard solution, a simple generic session_DTD has been written on the basis of the scenario_DTD. It describes the steps of the scenario with their respective input and output data. Depending on the desired usage of the data, appropriate XSL transformations should allow easy conversion of this generic representation of the retrieved data into desired more human readable documents.

The Xcollect application is a java application composed of two modules : the configuration module and the execution module. In the configuration module an interface allows the user to enter manually all the information specifying his scenario. Entered data are stored into an XML document according to the generic scenario_DTD. The execution module takes as input the XML scenario document, implements each step of the scenario and returns an XML document containing the retrieved data structured according to the generic Xcollect session_DTD.

METADATA RETRIEVAL FOR MANAGEMENT OF "HOMOLOGOUS" ANSWERS

A specific scenario (Xfunction) has been designed to retrieve functional annotations for given human genes. Source diversity implies multiple answers for the same query. Interpreting this diversity may be trivial for well-known genes but becomes crucial in the case of less-known genes presenting various putative functional assignments. Taken together and depending on where they come from, "homologous" answers may present identity, concordance, complementarity or discrepancy semantic relationships. Ideally, the user should be assisted in retrieving all possible answers, sorting them according to criteria of his choice - likely "quality criteria" about sources and/or data-, and checking for any inconsistency. Therefore an analysis step was required to clarify which metadata should be retrieved at the same time as the data themselves.

The exploration of available sources for functional annotations of genes led us to select the four sources listed in Table 1. The first two sources (SwissProt and PIR) are protein-oriented and contain entries corresponding to genes with known function. The other two sources (RefSeq and TIGR-HGI) can be queried to fetch annotations about less well-known genes. A great heterogeneity was observed in the designation of the fields containing function description in all these sources. Metadata reflecting source quality such as coverage, update frequency and existence of manual curation, have been valued for each source. Metadata associated with source entries, such as literature references, GO (The Gene Ontology Consortium, 2001) annotation, date of last modification, and status or tracking of the annotations, are then documented.

In addition to the fact that they contain functional annotations, the four sources have been selected on the basis of the existence of GO annotation, because this may guide reasoning about concordance or discrepancies between the answers. GO annotations are always traceable but entries without well-characterized gene product may lack such annotation.

The Xfunction scenario is schematized in Figure 1. Since it is not possible to query all the sources with a common identifier, the first step of the scenario involves querying either a unified query system such as SRS, or an integrated database such as GeneCards or GeneLynx, to retrieve all desired identifiers. In a second step the four selected sources are queried for functional annotations. Entry-associated metadata are retrieved in parallel with the data.

Table 1. Characteristics - in terms of quality metadata - of the sources used to retrieve information about gene function

Source	Functional Data	Source Metadata				Source Entry Metadata					
	Fields	Coverage		Update frequency	Manual Revision	Litterature references	GO annotation	Creation Date	Date of last modification		Annotation Tracking/Status
		Total (Dec 2003)	Human (Dec 2003)						Sequence	Annotation / Entry	
SwissProt	-COMMENTS : Function	141 681	10 404	year	yes	yes	yes [Ev[1]]	yes	yes	yes	no
PIR-PSD	- TITLE - ALTERNATE_ NAMES -FUNCTION : #description, #note	283 366	10 395	3 months	ongoing	yes	yes [Ev] under iProClass display	yes	yes	yes	no
RefSeq	- DEFINITION - ? COMMENT[2] : Summary	831 287	22 740	daily	ongoing	yes	yes [Ev]	no	yes[2]	yes	Status Key in"COMMENT" block
TIGR-HGI (THC reports)[3]	- Tentative Annotation - Similarity Search results	-	843 786	4 months	no	yes	yes [Ev]	no	no	no	no

[1] Ev : with evidence code such as Traceable Author Statement (see http://www.geneontology.org/GO.annotation.html)
[2] Only available for REVIEWED RefSeq entries
[3] THC reports are only available for genes without well-described transcript

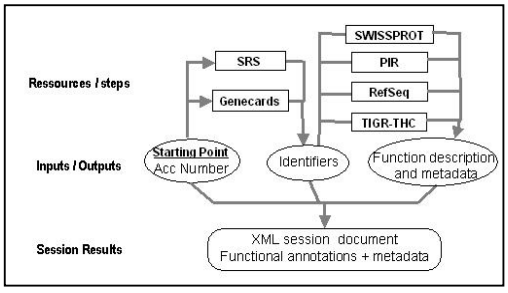

Figure 1 . Xfunction scenario. The resources queried at each step are indicated in boxes. Data (Inputs / Outputs) are circled.

Enactment of the resulting Xfunction scenario for a given gene thus provides a set of data and metadata, structured as an XML document so that it can be further processed. An additional module is being developed to exploit these data with regard to the metadata mentioned in Table 1. Thus, the biologist could define his own weighting scheme(s) of source and source-entry quality criteria that should be used to sort the answers as in Berti-Equille et al. (2001).

CONCLUSIONS AND PERSPECTIVES

Very recently, several large scale projects have been conducted to build integrative plateforms for bioinformatics resources. The most popular are probably myGRID (www.mygrid.org.uk ; Wroe et al., 2003) and BioMOBY (www.biomoby.org ; Wilkinson and Links, 2002). The problems they address include (i) the discovery of bioinformatics resources seen as web services, (ii) their composition into user-defined workflows (a lesson learned from these projects confirms that the biologists wish to be kept in the loop at this level) and (iii) the enactment of these worflows. Various standards are emerging for the description of workflows such as Wf-XML and XPDL by the Workflow Management Coalition (WfMC, http://www.wfmc.org/standards/standards.htm), or WSFL used by MyGrid for workflows involving web services. Once one of these standards will be stabilized, it will be interesting to adapt it to the description and enactment of our scenarios.

However a challenging problem remains unsolved : when the user is building his workflow, at each step, several relevant services are proposed to him and he must choose one of them. What happens when the different services are complementary (and possibly redundant or conflicting) to answer the question ? Beyond the gathering of homologous data and their sorting according to user preferences, it would be helpful to perform real semantic integration of these data, *i.e.*, to analyse their agreement level, their precision level, etc. This is the problem we want to address in future work on a basis compatible with the above-cited projects, *i.e.*, domain-based ontologies expressed in a web semantic language based on description logics.

REFERENCES

Berti-Equille,L. (2001) Integration of Biological Data and Quality-Driven Source Negotiation. In proc. 20th. Intl. Conference on Conceptual Modeling (ER2001), Lecture Notes in Computer Science, vol **2224**, 256-269, Yokohama, Japan.

Davidson,S.B., Crabtree,J., Brunk,B.P., Schug,J., Tannen,V., Overton,G.C. and Stoeckert, Jr.C.J. (2001) K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. IBM systems journal, **40**, 512-531.

Devignes,M.D., Schaaf,A. and Smaïl,M. (2002) Collecte et intégration de données biologiques hétérogènes sur le web – Xmap : application dans le domaine de la cartographie du génome humain. RSTI – ISI. Recherche et filtrage d'information, **7**, 45-61.

Eckman,B.A., Lacroix,Z. and Raschid,L. (2001) Optimized seamless integration of biomolecular data. IEEE symposium on Bio-Informatics and Biomedical Engineering (BIBE'2001), Washington DC, nov 2001, p.23-32.

Ernst,P., Glatting,K.-H. and Suhai,S. (2003) A task framework for the web interface. Bioinformatics,**19**, 278-282.

Goble,C., Paton,N., Stevens,R., Baker,N.G.P., Peim,M., Bechhofer,S. and Brass,A. (2001) Transparent Access to Multiple Bioinformatics Informations Sources. IBM systems journal, **40**, 532-551.

The Gene Ontology Consortium (2001) Creating the Gene Ontology Resource: Design and Implementation. Genome Research, **11**, 1425-1433.

Wilkinson,M. and Links,M. (2002) BioMOBY : An open source biological web services proposal. Briefings in Bioinformatics, **3**, 331-341.

Wroe,C., Stevens,R., Goble,C. and Greenwood,M. (2003) A suite of DAML+OIL ontologies to describe bioinformatics web services and data. International Journal of Cooperative Information Systems, **12**, 197-224.